Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency

Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer

Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency

Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
{firstname.lastname}@inria.fr

Abstract. The rise of virtual and augmented reality fuels an increased need for content suitable to these new technologies including 3D contents obtained from real scenes. We consider in this paper the problem of 3D shape reconstruction from multi-view RGB images. We investigate the ability of learning-based strategies to effectively benefit the reconstruction of arbitrary shapes with improved precision and robustness. We especially target real life performance capture, containing complex surface details that are difficult to recover with existing approaches. A key step in the multi-view reconstruction pipeline lies in the search for matching features between viewpoints in order to infer depth information. We propose to cast the matching on a 3D receptive field along viewing lines and to learn a multi-view photoconsistency measure for that purpose. The intuition is that deep networks have the ability to learn local photometric configurations in a broad way, even with respect to different orientations along various viewing lines of the same surface point. Our results demonstrate this ability, showing that a CNN, trained on a standard static dataset, can help recover surface details on dynamic scenes that are not perceived by traditional 2D feature based methods. Our evaluation also shows that our solution compares on par to state-of-the-art-reconstruction pipelines on standard evaluation datasets, while yielding significantly better results and generalization with realistic performance capture data.

Keywords: Multi View · Stereo Reconstruction · Learned Photoconsistency · Performance Capture · Volume Sweeping

1 Introduction

In this paper, we examine the problem of multi-view shape reconstruction of real-life performance sequences, in other words with realistic clothing, motions, and corresponding capture set assumptions. 3D reconstruction is a popular and mature field with numerous applications related to the ability to record and replay 3D dynamic scenes, as with for instance the growing domain of virtual and augmented reality. An essential and still improvable aspect in this matter, in particular with performance capture setups, is the fidelity and quality of the recovered shapes, our goal in this work.

Fig. 1. Challenging scene captured with a passive RGB multi-camera setup [1]. (*left*) one input image, (*center*) reconstructions obtained with classical 2D features [22], (*right*) proposed solution. Our results validate the key improvement of a CNN-learned disparity to MVS for performance capture scenarios. Results particularly improve in noisy, very low contrast and low textured regions such as the arm, the leg or even the black skirt folds, which can be better seen in a brightened version of the picture in Figure 8.

Multi-view stereo (MVS) based methods have attained a good level of quality with pipelines that typically comprise feature extraction, matching stages and 3D shape inference. Interestingly, very recent works have re-examined stereo and MVS by introducing features and similarity functions automatically inferred using deep learning. The main promise of this type of method, is to include better data-driven priors, either in 2D [40, 24, 41, 39] as improvement over classic 2D features, or in 3D to account for relative view placement and local or global shape priors [5, 17, 18]. These novel MVS methods have been tested on static scene benchmarks with promising results, offering the prospect of outperforming standard feature pipelines thanks to these data-aware feature measures.

Our main goal is to examine whether these improvements transfer to the more general and complex case of live performance capture, where a diverse set of additional difficulties arise. Typical challenges for these capture situations include smaller visual projection areas of objects of interest due to wider necessary fields of view for capturing motion; occlusion and self-occlusion of several subjects interacting together; lack of texture content typical of real-life subject appearance and clothing; or motion blur with fast moving subjects such as sport action scenes (see Figure 7). To the best of our knowledge, existing learning-based MVS schemes report results on static datasets such as DTU [16] or ShapeNet [4] but have not yet been demonstrated on performance capture data with the aforementioned typical issues.

With the aim to generalize to this type of data we propose a novel framework that takes advantage of recent learning methods while keeping the precision advantage of a per view depth map extraction, as applied in many successful MVS algorithms [28]. Our approach performs multi-view matching within local vol-

umetric units of inference. Contrary to previous methods, our volumetric unit is defined in a given view's own reference, so as to capture camera inherent 3D dependencies, specifically for the purpose of per-view decision. Instead of inferring occupancies, we infer disparity scores to ease training and to focus the method more on photometric configurations than local shape patterns. We sweep viewing rays with this volumetric receptive field, a process we coin *volume sweeping*, and embed the algorithm in a multi-view depth-map extraction and fusion pipeline followed by a geometric surface reconstruction. With this strategy, we are able to validate that CNN-based MVS outperforms classical MVS approaches in dynamic performance scenarios. We obtain high precision geometric results on complex sequences, outperforming both existing CNN-based and classic non-learning methods. We verify this improvement on available benchmarks with static objects. These results on diverse data situations are obtained using only a DTU subset as training data, which evidences the generalization capabilities of our network.

2 Related Work

Multi-view stereo reconstruction is a longstanding active vision problem [32]. Initially applied on static scenes, the extension to performance capture of dynamic scenes has become increasingly popular. Stereo and MVS-based approaches are a modality of choice for high fidelity capture applications [12, 34, 13, 29, 16, 27, 31], possibly complementing other strategies such as depth-based reconstruction [28, 15, 6, 10] by addressing shortcomings that include limited range, sensitivity to high contrast lighting, and interference when increasing the number of viewpoints.

While considering various shape representations, for instance point clouds [12], fused depth maps [25], meshes [33, 21], or volumetric discretizations [20, 8, 38], most MVS methods infer 3D shape information by relying on the photoconsistency principle that rays observing the same scene point should convey similar photometric information.

In its simplest form, such similarity can be measured by considering projected color variances among views, as used in early works [20] with limited robustness. In stereo and short baseline situations, simple normalized forms of 2D window correlation are sufficient to characterize similarity under simple lighting and constrast changes, using *e.g.* ZNCC, SSD, SHD. For broader geometric and photometric resilience, various features based on scale-invariant gradient characterizations [23, 2, 26] have been designed, some specialized for the dense matching required for the MVS problem [36]. More recently, image features have been successfully applied to moving sequences in *e.g.* [27, 22]. Generally, MVS methods characterize photoconsistency either with a symmetric, viewpoint agnostic, combination of all pairwise similarities [30], or with a per image depth map determination through sweeping strategies [7, 25]. Our approach employs also a sweeping strategy, which proves generally simpler and still significantly

more robust to occlusion than view-agnostic methods, an issue that quite often occurs in practice with multiple moving shapes or through limb self-occlusion.

While classic MVS approaches have been generally successful, recent works aimed at learning stereo photoconsistency have underlined that additional priors and more subtle variability co-dependencies are still discoverable in real world data. Several works leverage this by learning how to match 2D patch pairs for short baseline stereo, letting deep networks infer what features are relevant [40, 24, 41, 39]. Very recent works extend this principle to wide baseline MVS, with symmetric combination of 2D learned features [14].

The common limitation of such methods with 2D receptive fields is the difficulty to correctly capture 3D correlations with hence both false positive and false negative correlations arising from the 2D projection. Consequently, a number of learned MVS methods resort to full volumetric 3D receptive fields instead, to broaden the capability to any form of data 3D correlation [5, 17, 18]. While casting correlations in 3D as well, our approach proposes several key differences: our volumetric receptive field is a back-projected image region, similar to some binocular stereo [19] or image-based rendering [11] works, where the latter only uses the grid as proxy without explicitely extracting 3D information. This enables a sweeping search strategy along viewing rays, which proves a robust search strategy as plane sweeping in stereo reconstruction. This scheme also avoids decorrelating camera resolution and 3D receptive field resolution, as with *e.g.* voxels, the volumetric receptive field being defined as a backprojection along pixel rays. Additionally, this volumetric receptive field learns local pairwise correlations, a lower level and easier task than learning occupancy grid patterns. Our evaluation on practical performance capture scenes, beyond traditional static datasets, validates the benefit of such a learning strategy over traditional approaches.

3 Method Overview

As for many recent multi-view stereo reconstruction methods, ours estimates per camera depth maps, followed by depth fusion, allowing therefore each camera to provide local details on the observed surface with local estimations. We take this strategy a step further by replacing the traditional photoconsistency measure used to estimate depths with a learned version. This version is based on CNNs and exploits their ability to learn local photometric configurations near surfaces observed from multiple viewpoints. As depicted in Figure 2, our approach takes as input a set of calibrated images and outputs a 3D mesh obtained by fusing depth maps. Depths along pixel viewing rays are obtained using a volume sweeping strategy that samples multi-view photoconsistency along rays and identifies the maxima. For a point along a viewing ray, the photoconsistency is estimated using a discretized 3D volumetric patch around that point. In such a 3D patch, at each point within, color information from the primary camera ray incident to that point is paired to the color information of the incident ray of another camera. We collect these paired color volumes for every other camera

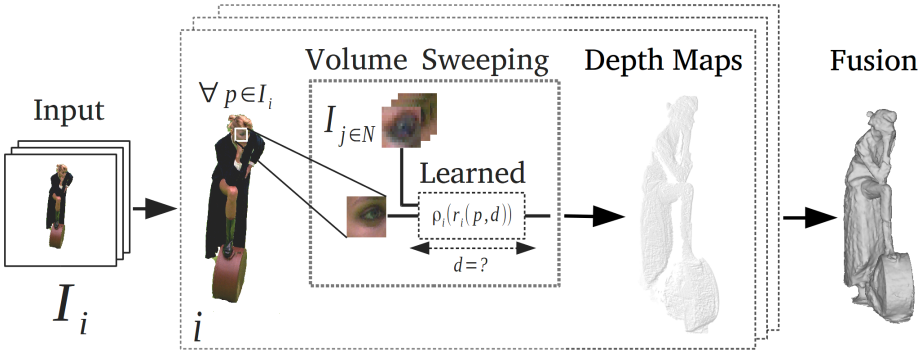

Fig. 2. Method pipeline and notations.

than the primary. A trained CNN is used to recognize the photoconsistent configurations given pairs of color samples within the 3D patch. The key aspects of this strategy are:

- The per camera approach, which, by construction, samples the photoconsistency at a given location as captured and thus enables more local details to be revealed compared to global approaches, as shown in Figure 8.
- The 3D receptive field for the photoconsistency evaluation, which resolves some 2D projection ambiguities that hindered 2D based strategies.
- The learning based strategy using a convolutional neural network, which outperforms traditional photometric features when evaluating the photoconsistency in dynamic captured scenes, as demonstrated by our experiments.

The following sections focus on our main contributions, namely the 3D volume sampling and the learning based approach for the photoconsistency evaluation. Note that for the final step, without loss of generality, we use the TSDF to fuse depth information and [22] to get a 3D mesh from the fused depths.

4 Depth Map Estimation by Volume Sweeping

Our reconstruction approach takes as input N images $\{I_i\}_{i=1}^N$, along with their projection operators $\{\pi_i\}_{i=1}^N$, and computes depth maps, for the input images, that are subsequently fused into a 3D implicit form. This section explains how these maps are estimated. Given a pixel p in an input image i, the problem is therefore to find the depth d along its viewing ray of its intersection with the observed surface. The point along the ray of pixel p at depth d is noted $r_i(p, d)$. Our approach searches along viewing rays using a likelihood function for a point to be on the surface given the input color pairs in the evaluation volume. In contrast to traditional methods that consider hand-crafted photoconsistency measures, we learn this function from multiview datasets with ground-truth surfaces. To this purpose we build a convolutional neural network which, given a reference camera i and a query point $x \in \mathbb{R}^3$, maps a local volume of color pair samples around

x to a scalar photoconsistency score $\rho_i(x) \in [0..1]$. The photoconsistency score accounts in practice for color information from camera i at native resolution, and for other camera colors and their relative orientation implicitly encoded in the volume color pair construction. These important features allow our method to adapt to specific ray incidences. Its intentionally asymmetric nature also allows subsequent inferences to automatically build visibility decisions, *e.g.* deciding for occlusion when the primary camera i's color is not confirmed by other view's colors. This would not have been possible with a symmetric function such as [14].

We thus cast the photoconsistency estimation as a binary classification problem from these color pairs around x, with respect to the reference image i and the other images. In the following, we first provide details about the 3D sampling regions before describing the CNN architecture used for the classification and its training. We then explain the volume sweeping strategy that is subsequently applied to find depths along rays.

4.1 Volume Sampling

In order to estimate photoconsistency along a viewing ray, a 3D sampling region is moved along that ray at regular distances. Within this region, pairs of colors backprojected from the images are sampled. Each pair contains a color from the reference image and its corresponding color in another image. Samples within the 3D region are taken at regular depths along viewing rays in the reference image (see Figure 3). The corresponding volume is a truncated pyramid that projects onto a 2D region of constant and given dimension in the reference image. This allows the 3D sampling to adapt to the camera perception properties, *e.g.* resolution and focal length.

More precisely, consider the back-projection $r_i(p, d)$ at depth d of pixel p from the reference image i. The k^3 input sample grid used to compare pairs of colors from images $\{i, j\}_{j \neq i}$ is then the set of back-projected pixels in a k^2 window centered on p, regularly sampled from depth $d - k\lambda/2$ to $d + k\lambda/2$, with λ chosen s.t. spacing in the depth direction is equal to inter-pixel distance from the reference camera at that depth. Every sample contains the reference color of the originating pixel in image i and the color of the point projected on camera j.

Volume sampling is always performed with the same orientation and ordering with respect to the reference camera. Convolutions are thus consistently oriented relative to the camera depth direction.

Volume Size In our experiments and with no loss of generality, $k = 8$. Our strategy is to learn pairwise photoconsistent configurations along rays, in order to detect the surface presence. This is in contrast with previous works that try to infer directly shape within regular voxel grids, *e.g.* [17] with 32^3 or 64^3 grids. By considering the surface detection problem alone, and letting the subsequent step of fusion integrate depth in a robust and consistent way, we simplify the problem and require little spatial coherence, hence allowing for small grids.

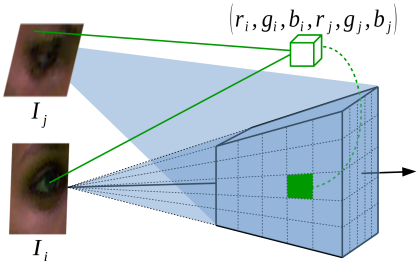

Fig. 3. The 3D volume used to estimate photoconsistency along rays from the reference image i. k^3 samples within the volume are regularly distributed along viewing rays and contain color pairs as back-projected from images i and j. At a given depth along a ray from i each image $j \neq i$ defines a pairwise comparison volume.

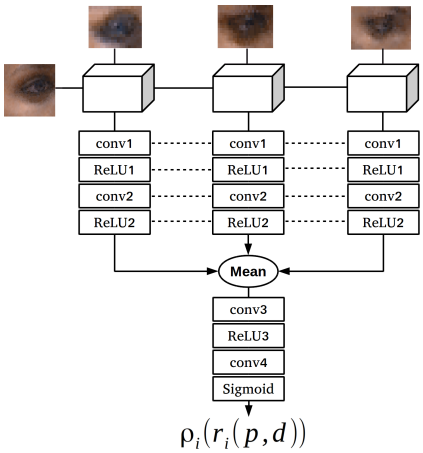

Fig. 4. CNN architecture. Each cube is a pairwise comparison volume with k^3 samples that contain 6 valued vectors of RGB pairs and over which 3D convolutions are applied. The output score $\rho_i(r_i(p,d)) \in [0..1]$ encodes the photoconsistency at depth d along the ray from pixel p in image i.

4.2 Multi-View Neural Network

As explained in the previous section, at a given point x along a viewing ray we are given $N-1$ volumes colored by pairs of views, *i.e.* $(N-1) \times k^3$ pairs of colors, and we want to detect whether the surface is going through x. To this aim, we build siamese encoders similarly to [14], with however 3D volumes instead of 2D patches. Each encoder builds a feature given a pairwise volume. These features are then averaged and fed into a final decision layer. Weight sharing and averaging are chosen to achieve camera order invariance.

The network is depicted in Figure 4. The inputs are $N-1$ colored volumes of size $k^3 \times 6$ where RGB pairs are concatenated at each sample within the volume. Convolutions are performed in 3D over the 6 valued vectors of RGB pairs. The first layers (encoders) of the network process every volume in parallel, with shared weights. Every encoder is a sequence of two convolutions followed by non-linearities, and max-pooling with stride. Both convolutional layers consist of respectively 16 and 32 filters of kernel $4 \times 4 \times 4$, followed by a Rectified Linear Unit (ReLU) and a max-pooling with kernel $2 \times 2 \times 2$ with stride 2. We then average the obtained $2 \times 2 \times 2 \times 32$ features and feed the result to a 128 filter $1 \times 1 \times 1$ convolutional layer, followed by a ReLU and a final $1 \times 1 \times 1$ decision layer, for a total of $72K$ parameters . The network provides a score $\rho_i(r_i(p,d)) \in [0..1]$ for the photoconsistency at depth d along the ray from pixel p in image i.

We experimented with this network using different configurations. In particular, instead of averaging pairwise comparison features, we tried max-pooling

which did not yield better results. Compared to the volumetric solution proposed by [17], the number of parameters is an order of magnitude less. As mentioned earlier, we believe that photoconsistency is a local property that requires less spatial coherence than shape properties.

4.3 Network Training

The network was implemented using TensorFlow and trained from scratch using the DTU Robot Image Dataset [16], which provides multiview data equipped with *ground-truth* surfaces that present an accuracy up to $0.5mm$. From this dataset 11 million k^3 sample volumes were generated, from which we randomly chose 80 percent for training, and the remaining part for evaluation. Both positive and negative samples were equally generated by randomly sampling volumes up to $20cm$ away from ground truth points, where a volume is considered as positive when it contains at least μ ground truth points. In theory, the network could be trained with any number of camera pairs, however, in practice, we randomly choose from one up to 40 pairs. Training was performed with the binary cross entropy function as loss. Model weights are optimized by performing a Stochastic Gradient Descent, using Adaptive Moment Estimation on $560,000$ iterations with batch size of 50 comparisons, and with a random number of compared cameras (from 2 up to 40). Since our sampling grids are relatively small and camera dependent, we are able to generate enough sample variability for training, without the need for data augmentation.

4.4 Volume Sweeping

In order to estimate the depth along viewing rays, our volumetric solution is integrated in an existing standard plane sweeping algorithm, replacing the plane with a volume and computing the N-way photoconsistency score using our network. For every camera, we sample therefore along viewing rays, test possible depth values, and choose the most photoconsistent candidate with respect to the network score. In practice, a reference view i is only compared to the cameras such that $\cos(\theta_{ij}) > 0.5$, where θ_{ij} is the angle between the optical axes of camera i and j. Then, we sample rays from camera i through every pixel p and build colored volumes at every candidate depth. We define the estimated depth d_i^p as:

$$d_i^p = \underset{d \in [d_{min}, d_{max}]}{\operatorname{argmax}} \; (\rho_i(r_i(p, d))), \tag{1}$$

where $\rho_i(r_i(p, d))$ is the consistency measure along the ray from p in image i, as estimated by the network, and $[d_{min}, d_{max}]$ defines the range of search that can be limited using for instance the visual hull when available. Depths for all pixels and from all images are further fused using a volumetric truncated signed distance function [9].

5 Results

We perform various evaluations to verify and quantify the benefit of our learned multi-view similarity. First, we study different classifiers performances, with an emphasis on comparing planar or volumetric receptive fields. We next apply our approach in the static case using the [16] benchmark and compare it to state-of-the-art MVS methods, both classic and learning based. Finally, we build experiments to test the main claim of improvement with real life performance data. To this goal we use several captured dynamic sequences which exhibit typical difficulties of such data, with very significant qualitative improvements compared to the state-of-the-art approaches [17], and [22].

5.1 Surface Detection

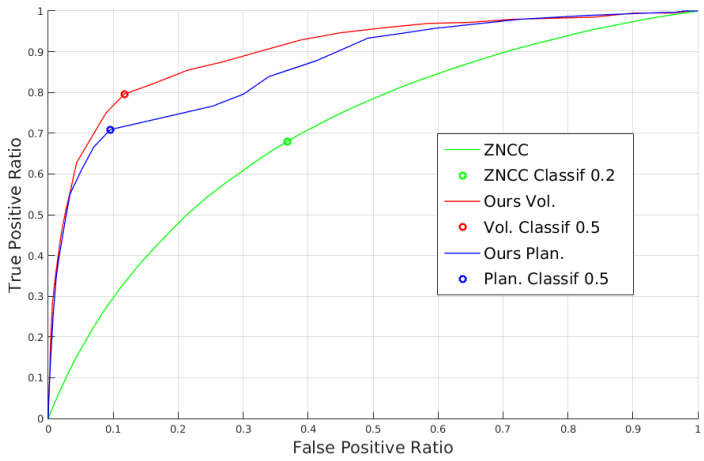

Fig. 5. ROC Curves of three different classifiers, ZNCC, planar and volumetric receptive fields, on the DTU Dataset [16]. Circles represent thresholds that optimize sensitivity + specificity with the values 0.2, 0.5 and 0.5 respectively.

Surface detection along viewing rays can be formulated as a binary classification problem. In order to assess the benefit of our volumetric strategy, we compare performances of classifiers based on various receptive fields.

1. Deterministic Zero-Mean Normalized Cross Correlation (*ZNCC*): ZNCC is applied over the samples within the volumetric receptive field.
2. Learning (CNN) with a planar receptive field: a planar equivalent of our volumetric solution, with the same architecture and number of weights, in a fronto-facing plane sweeping fashion.

3. Learning (CNN) with a volumetric receptive field: our solution described in the previous sections.

To speed up computations, we limit the search along a viewing ray to $5mm$ around a coarse depth estimation based on image descriptors [35]. Depths are sampled every $0.5mm$. As a post processing step, we simply add a soft bilateral filter, similarly to [14], accounting for color, spatial neighborhood, and probability of the detection. Figure 5 shows, with the classifiers' ROC curves, that the most accurate results are obtained with a volumetric receptive field and learning. Intuitively, a volumetric sampling region better accounts for the local non-planar geometry of the surface than planar sampling regions. This graph also emphasizes the significantly higher discriminative ability of learned correlations compared to deterministic ones.

We also evaluate the robustness to baseline variability by testing classification with more further apart cameras. Table 2 shows the accuracy of the classifiers with a varying number of cameras and for the optimal threshold values in Figure 5 . As already noticed in the litterature, *e.g.* [12, 29], a planar receptive field gives better results with a narrow baseline and the accuracy consistently decreases when the inter-camera space grows with additional cameras. In contrast the classifier based on a volumetric receptive field exhibits more robustness to the variety in the camera baselines. This appears to be an advantage with large multi-camera setup as it enables more cameras to contribute and hence reduces occlusion issues.

5.2 Quantitative Evaluation

In this section, we compare our solution to various state-of-the-art methods using the DTU Robot Image Dataset [16]. We use the standard accuracy and completeness metrics to quantify the quality of the estimated surface. We compare to Furukawa et al. [12], Campbell et al. [3] and Tola et al. [36], as well as to additional learning-based results from Ji et al. [17] and Hartmann et al. [14]. To conduct a fair comparison with [14], which is a patch based approach building a depthmap with a network comparable to ours, we use the result of our volume sweeping approach on only one depth map.

Reconstructions results are depicted in table 1. We obtain quality on par with other methods, with a median accuracy and completeness in the range of the ground truth accuracy that we measured around $0.5mm$. It should be noticed that the best accuracy is obtained by Tola et al. [37] which tend to favor accuracy over completeness whereas Campbell et al. [3], in a symetric manner, tend to favor completeness over accuracy. We obtain more balanced results on the 2 criteria, similarly to the widely used approach by Furukawa et al. [12], with however better performances. We also outperform the recent learning based method Surfacenet [17] on most measures in this experiment.

Compared to Hartmann et al. [14], and under similar experimental conditions, our approach obtains better results with 2 orders of magnitude less parameters, thereby confirming the benefit of volumetric receptive fields over

planar ones. Compared to Surfacenet [17] (cube size $64 \times 64 \times 64$, sample step $0.4mm$) we obtain reconstructions of slightly better quality with an order of magnitude less parameters.

Table 1. Reconstruction accuracy and completeness (in mm).

	Acc.		Compl.	
Measure	Mean	Med.	Mean	Med.
Tola et al. [37]	**0.448**	**0.205**	0.754	0.425
Furukawa et al. [12]	0.678	0.325	0.597	0.375
Campbell et al. [3]	1.286	0.532	**0.279**	**0.155**
Ji et al. [17]	0.530	0.260	0.892	0.254
Ours (*fused*)	0.490	0.220	0.532	0.296
Hartmann et al. [14]	1.563	0.496	1.540	0.710
Ours (*depthmap*)	**0.599**	**0.272**	**1.037**	**0.387**

Table 2. Classifier accuracy (%).

Camera #	5	20	49
ZNCC	64.98	65.46	65.58
Ours Plan.	80.67	77.87	75.92
Ours Vol.	**82.95**	**84.84**	**83.45**

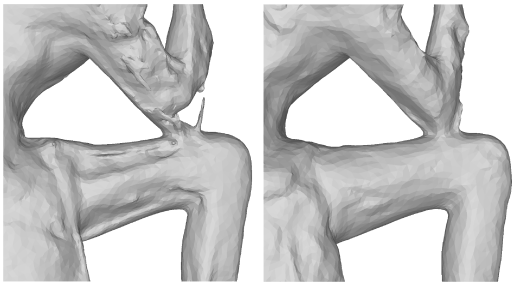

Fig. 6. Close up view of the arm region in Figure 1. (*Left*) Results from [22], (*right*) our reconstruction

5.3 Qualitative Evaluation and Generalization

One of our main goals is to verify whether a learning based strategy generalizes to the performance capture scenario and how it compares to state-of-the-art deterministic approaches in this case. To this purpose, we perform reconstructions of dynamic RGB sequences captured by a setup largely different from the training one, *i.e.* a hemispherical setup with 68 cameras of $4M$ resolution with various focal lengths, as provided in [22] along with reconstructions obtained with a deterministic approach. In this scenario, standard MVS assumptions are often violated, *e.g.* specular surfaces, motion blur and occlusions, challenging therefore the reconstruction methods.

Fig. 7. (*Top*) input images, (*middle*) result with [22], (*bottom*) result with our method. Motion blur and low contrast are visible in the input images . Best viewed magnified.

We adapted our volume sweeping algorithm to limit depth search, along viewing rays, inside visual hulls. No other modification was applied, in particular the network previously trained was kept as such without any fine tuning. Figure 1 shows a reconstruction using our method compared to [22], which is a patch based sweeping method using traditional image features and specifically designed for this scenario. Even though [22] performs well in contrasted regions, the patch based descriptors reach their limits in image regions with low contrast or low resolution. Figure 6 and 7 give such examples. They show that our solution helps recover finer surface details, while strongly decreasing noise in low contrast regions. The results obtained also demonstrate strong improvements in surface details, such as dress folds, that were undetected by the deterministic approach. In addition, they demonstrate lower levels of noise, particularly in self-occluded regions, and more robustness to motion blur as with the toes or tongue-in-cheek details that appear in Figure 7-bottom.

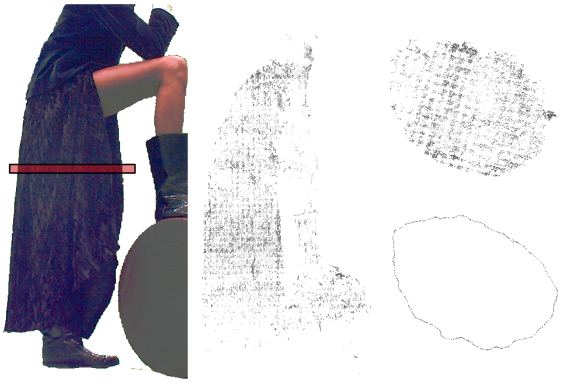

Fig. 8. Qualitative comparison with [17]. (*Left*) input image with the horizontal section in *red*, (*middle*) point cloud with [17], (*right-top*) point cloud horizontal section with [17] (*right-bottom*) point cloud horizontal section with our approach.

We also compared with a recent learning based approach [17] using the code available online (see Figure 8). Reconstructions with this approach were limited to a tight bounding box and different values for the volume sampling step were tested. The best results were obtained with a $2mm$ step. To conduct a fair comparison with our method, all points falling outside the visual hull were removed from the reconstruction. In this scenario, the point cloud obtained using [17] appeared to be very noisy and incomplete (see Figure 8-middle), plaguing the subsequent surface extraction step. Figure 8-left also shows a horizontal section of the model in a poorly contrasted image region of the dress. The global strategy used in [17] wrongly reconstruct many surface points inside the shape volume (top figure), as a result of the ambiguous appearance of the dress. In contrast, our approach (bottom figure) correctly identifies surface points by maximizing learned correlations along viewing rays.

Fig. 9. (*Left*) 3 input images, (*middle*) plane based classifier, (*right*) volumetric classifier. The face is highly occluded (*left*) yielding noisier and less accurate reconstructions when using a planar receptive field, whereas the volume counterpart yields smoother and more accurate details.

The final qualitative experiment studies the impact of a volumetric receptive field compared to the equivalent planar one (see sec. 5.1) in figure 9. The volume allows a sharp reconstruction of finer details of the belt, where a plane cannot handle finer geometry details. A video demonstrating results on dynamic sequences is available online: https://hal.archives-ouvertes.fr/hal-01849286.

6 Conclusion

We presented a learning framework for surface reconstruction in passive multi-view scenarios. Our solution consists in a N-view volume sweeping, trained on static scenes from a small scale dataset equipped with ground truth. Thanks to this new model, we validate the improvement of CNN-learned MVS similarity in the case of complex moving sequence captures, with significant challenges typical of these datasets such as low light areas and low texture content and perceived resolution. This result is achieved with an order of magnitude less training parameters than previous comparable learned MVS works, showing significant network generalization from a training performed only on static DTU inputs, and fully leverages the high quality ground truth now available with these datasets. Our method achieved significantly improved detail recovery and noise reduction in complex real life scenarios, outperforming all existing approaches in this case, and consequently offers very interesting prospects for even more challenging capture scenarios or even better ground truth datasets in the future.

Acknowledgements

Funded by France National Research grant ANR-14-CE24-0030 ACHMOV. Images 1-2-6-8 courtesy of Anja Rubik.

References

1. Kinovis inria platform. https://kinovis.inria.fr/inria-platform/
2. Bay, H., Tuytelaars, T., Gool, L.J.V.: SURF: speeded up robust features. In: ECCV (2006)
3. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: ECCV (2008)
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR] (2015)
5. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)
6. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A.G., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. (2015)
7. Collins, R.T.: A space-sweep approach to true multi-image matching. In: CVPR (1996)
8. Cremers, D., Kolev, K.: Multiview stereo and silhouette consistency via convex functionals over convex domains. IEEE Trans. Pattern Anal. Mach. Intell. (2011)
9. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: SIGGRAPH (1996)
10. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S.: Fusion4d: Real-time performance capture of challenging scenes. ACM Trans. Graph. (2016)
11. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. In: CVPR (2016)
12. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. In: CVPR (2007)
13. Gall, J., Stoll, C., Aguiar, E.D., Theobalt, C., Rosenhahn, B., peter Seidel, H.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR (2009)
14. Hartmann, W., Galliani, S., Havlena, M., Van Gool, L., Schindler, K.: Learned multi-patch similarity. In: ICCV (2017)
15. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: ECCV (2016)
16. Jensen, R.R., Dahl, A.L., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR (2014)
17. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In: ICCV (2017)
18. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In: NIPS (2017)
19. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV (2017)
20. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. IJCV (2000)
21. Labatut, P., Pons, J., Keriven, R.: Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: ICCV (2007)

22. Leroy, V., Franco, J.S., Boyer, E.: Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In: ICCV (2017)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
24. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: CVPR (2016)
25. Merrell, P., Akbarzadeh, A., Wang, L., michael Frahm, J., Nistr, R.Y.D.: Real-time visibility-based fusion of depth maps. In: CVPR (2007)
26. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: CVPR (2003)
27. Mustafa, A., Kim, H., Guillemaut, J., Hilton, A.: Temporally coherent 4d reconstruction of complex dynamic scenes. In: CVPR (2016)
28. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: CVPR (2015)
29. Oswald, M.R., Cremers, D.: A convex relaxation approach to space time multi-view 3d reconstruction. In: ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD) (2013)
30. Pons, J.P., Keriven, R., Faugeras, O.: Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. IJCV (2007)
31. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
32. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
33. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Comput. Graph. Appl. (2007)
34. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. CVPR (2008)
35. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: CVPR (2008)
36. Tola, E., Lepetit, V., Fua, P.: DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. (2010)
37. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach. Vis. Appl. (2012)
38. Ulusoy, A.O., Geiger, A., Black, M.J.: Towards probabilistic volumetric reconstruction using ray potentials. In: 3DV (2015)
39. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: CVPR (2017)
40. Žbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. (2016)
41. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: CVPR (2015)